FG-CXR: A Radiologist-Aligned Gaze Dataset for Enhancing Interpretability in Chest X-Ray Report Generation

Trong Thang Pham*[1], Ngoc-Vuong Ho[1], Nhat-Tan Bui[1], Thinh Phan[1], Patel Brijesh[3], Donald Adjeroh[3], Gianfranco Doretto[3], Anh Nguyen[4], Carol C. Wu[5], Hien Nguyen[2], and Ngan Le[1]

[1]University of Arkansas, Fayetteville, AR, USA
[2]University of Houston, Houston, Texas, USA
[3]West Virginia University, Morgantown, West Virginia, USA
[4]University of Liverpool, UK
[5]MD Anderson Cancer Center, Houston, Texas, USA
*Corresponding author: tp030@uark.edu

Abstract. Developing an interpretable system for generating reports in chest X-ray (CXR) analysis is becoming increasingly crucial in Computer-aided Diagnosis (CAD) systems, enabling radiologists to comprehend the decisions made by these systems. Despite the growth of diverse datasets and methods focusing on report generation, there remains a notable gap in how closely these models's generated reports align with the interpretations of real radiologists. In this study, we tackle this challenge by initially introducing *Fine-Grained CXR* (FG-CXR) dataset, which provides fine-grained paired information between the captions generated by radiologists and the corresponding gaze attention heatmaps for each anatomy. Unlike existing datasets that include a raw sequence of gaze alongside a report, with significant misalignment between gaze location and report content, our FG-CXR dataset offers a more grained alignment between gaze attention and diagnosis transcript. Furthermore, our analysis reveals that simply applying black-box image captioning methods to generate reports cannot adequately explain which information in CXR is utilized and how long needs to attend to accurately generate reports. Consequently, we propose a novel *explainable radiologist's attention generator* network (Gen-XAI) that mimics the diagnosis process of radiologists, explicitly constraining its output to closely align with both radiologist's gaze attention and transcript. Finally, we perform extensive experiments to illustrate the effectiveness of our method. Our datasets and checkpoint is available at https://github.com/UARK-AICV/FG-CXR.

Keywords: Chest X-ray · CXR Dataset · Intepretability · Deep Learning · Report Generation · Medical Imaging

1 Introduction

Chest X-rays (CXRs) are commonly used for both screening and diagnostic purposes, resulting in a substantial daily workload. Additionally, the current

shortage of trained radiologists in many healthcare systems highlights the need for automated radiology report generation to help reduce radiologists' workloads [51]. The success of Deep Learning [5, 21, 22, 30–32, 35, 37, 46, 48] has led people to pursue its application in the medical domain [1, 29]. However, most existing methods lack explainability, which is a major reason for their limited adoption. In the safety-critical medical field, a highly accurate but opaque report generation system may not be adopted if the reasoning behind the generated report is not transparent and explainable [11,12,27]. Therefore, creating and using an interpretable system should be prefer to black-box system [37].

In the examination process, radiologists carefully examine every anatomy of CXRs and report their findings. Inspired by this process, we hypothesize that understanding pixel importance and gaze patterns can improve AI model explainability and accuracy in CXR diagnosis. However, the use of radiologist gaze-derived heatmaps in generating descriptive reports during CXR diagnosis remains underexplored. Recently, Tanida et al. [43] address this challenge by introducing an interpretable system that uses bounding boxes, which lack detail. In contrast, Pham et al. [34] propose a diagnosis system directly supervised by gaze attention. However, this system is limited as it can only predict whether an anatomical region is abnormal, requiring users to identify the specific findings themselves, which can be impractical.

To address the aforementioned weaknesses, we introduce Gen-XAI pipeline, shown in Figure 1. Gen-XAI mimics how radiologists perceive images by decoding radiologist's gaze attention with the Gaze Attention Predictor and then explaining its observations through the Report Generator. The Gaze Attention Predictor focuses on learning the regions of interest based on radiologists' gaze attentions, ensuring that the system captures the critical areas that a radiologist would typically examine. The Report Generator then uses this information to produce an accurate radiology report, which is visually grounded with the anatomical gaze attention, enhancing the transparency and explainability of the diagnostic process.

Existing gaze datasets [1, 17] provide raw gaze sequences along with reports for each patient. However, radiologists typically observe before diagnosing, leading to a misalignment between the gaze location and the report at the same timestamp, as illustrated in Figure 2. Therefore, a cleaner dataset is needed to evaluate this pipeline effectively. To address this, we curate a new dataset that provides gaze sequences aligned with anatomical attention heatmaps. By aligning gaze sequences with attention heatmaps, we ensure that the generated reports are not only accurate but also provide insights into the reasoning process behind each diagnosis. Our main contributions are summarized as follows:

- We introduce FG-CXR, a curated dataset that provides anatomical segmentation, gaze attention heatmaps annotated by radiologists, and radiology reports that are aligned with the gaze attention heatmaps.
- We propose a novel interpretable baseline Gen-XAI to efficiently generate radiology reports with meaningful attention heatmap.

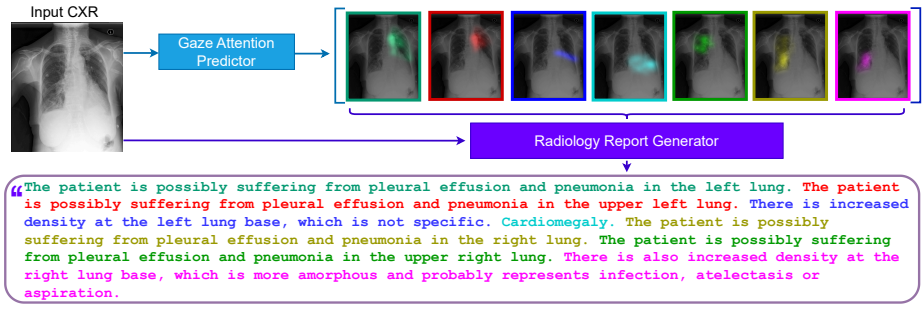


Fig. 1: An overview of our interpretable Gen-XAI framework, generating a diagnosis report and its corresponding visual attention for each diagnosis in the report.

2 Related works

2.1 Interpretable Deep Learning.

In high-stakes medical settings, understanding the decision-making process is crucial [38]. A direction to enhance interpretability is to design an architecture that can learn concepts [20, 34]. *In our paper, we follow the interpretable approach [38] by learning radiologists' intentions (gaze attention) across anatomical parts. However, unlike previous works [28, 34, 40] focusing on classification, we address the less explored task: the model must explain observations via report generation based on inferred intentions.*

2.2 Interpretable-oriented Datasets.

Creating datasets with annotated abnormality localization traditionally involves manual curation, but this is labor-intensive and often yields limited coverage, typically 1-2 labels [10, 42]. Recent efforts address this by providing datasets with anatomy labels in reports. For instance, the Chest ImaGenome dataset [51] is a dataset containing localized annotations (bounding box), with corresponding reports for the associated CXR images. However, existing datasets lack the granularity needed, e.g. gaze, to develop models that mimic real radiologist diagnoses. *In contrast, our dataset enhances detail by mapping reports to 7 anatomical locations using radiologist attention heatmaps, providing deeper insights into the diagnostic process.*

2.3 Radiology Report Generation.

Early approaches [15,24] in report generation leveraged CNN-RNN architectures or transformer inspired by general image captioning. However, medical report generation differs from image captioning [43] due to varying lengths, complexities, and biases in normal samples. To address these challenges, some models align visual features with disease tags [53], while others incorporate medical

Table 1: Overview of CXR datasets. A coarse report describes the whole image. A fine-grained report is a report associated with individual anatomies.

Gaze	Datasets	Annotation		#Samples	NLP Reports
		Information	Method		
✗	SIIM-ACR Pneumothorax Segmentation [10]	Segmentation	Manual + augmented	12,047	No
	RSNA Pneumonia Detection Challenge [42]	Bounding Boxes	Manual	30,000	No
	NIH CXR dataset [49]	Entire CXR	Automated	112,120	No
	PLCO [45]	Entire CXR	Automated	236,000	No
	Stanford CheXpert [13]	Entire CXR	Automated	224,316	No
	Montgomery County Chest X-ray [14]	Segmentation	Manual	138	No
	Shenzen Hospital Chest X-ray [14]	Segmentation	Manual	662	No
	Indiana University Chest X-ray Collection [8]	Entire CXR	Automated	3,813	Coarse
	MIMIC-CXR [16]	Entire CXR	Automated	377,110	Coarse
	Dutta [7]	Entire CXR	Manual	2,000	Coarse
	PadChest [2]	Entire CXR	Manual + automated	160,868	Coarse
	Chest ImaGenome [51]	Bounding Boxes	Automated	242,072	Fine-Grained
✓	REFLACX [1]	Gaze	Automated	3,000	Coarse
	EGD [17]	Segmentation + Gaze	Automated	1,000	Coarse
	Our FG-CXR	Atanomies Localization Gaze Attention Heatmap Gaze Sequence	Semi-automated Automated Automated	2,951	Fine-Grained

knowledge graphs [25]. Notably, RGRG [43] tackles interpretability by outlining abnormal regions and generating captions about them, but it lacks precision in specifying abnormality areas within bounding boxes. *In contrast, our method simulates radiologists' focus on important regions and generates insights based on them.*

Our FG-CXR dataset closely simulates radiologists' real-life diagnostic process by providing detailed annotations, including anatomical localization, gaze attention, and corresponding medical reports. A comparison between our FG-CXR with the existing CXR datasets is given in Table 1, while a visualization of the comparison of gaze-based annotations is shown in Fig. 2.

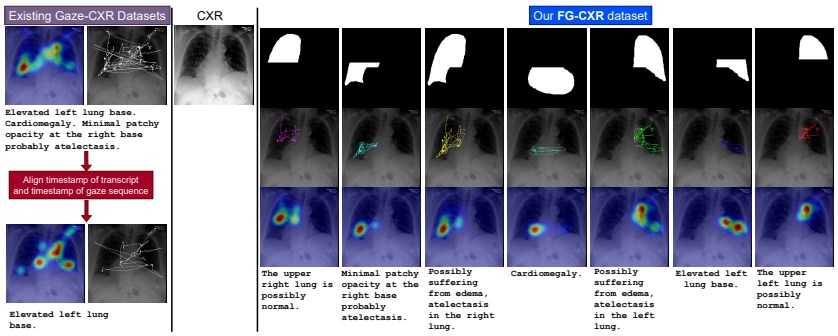


Fig. 2: Annotation comparison between prior gaze-CXR datasets (left) which face challenges in aligning gaze location with textual description and our FG-CXR (right), given a CXR (middle).

Table 2: Keywords for Anatomical Regions. The keywords are chosen by listing all sentences from the radiology reports and picked based on their meaning.

Anatomical Region	Keywords
Heart	cardiomegaly, enlarged and chest, heart, cardiac, mediastinum
Left Lung	left
Right Lung	right
Upper Left Lung	upper and left, apex and left, mid and left, apical and left, top and left
Upper Right Lung	upper and right, apex and right, mid and right, apical and right, top and right
Lower Left Lung	lower and left, base and left, bottom and left
Lower Right Lung	lower and right, base and right, bottom and right

3 Dataset: Fine-Grained CXR (FG-CXR)

3.1 Anatomy Localization.

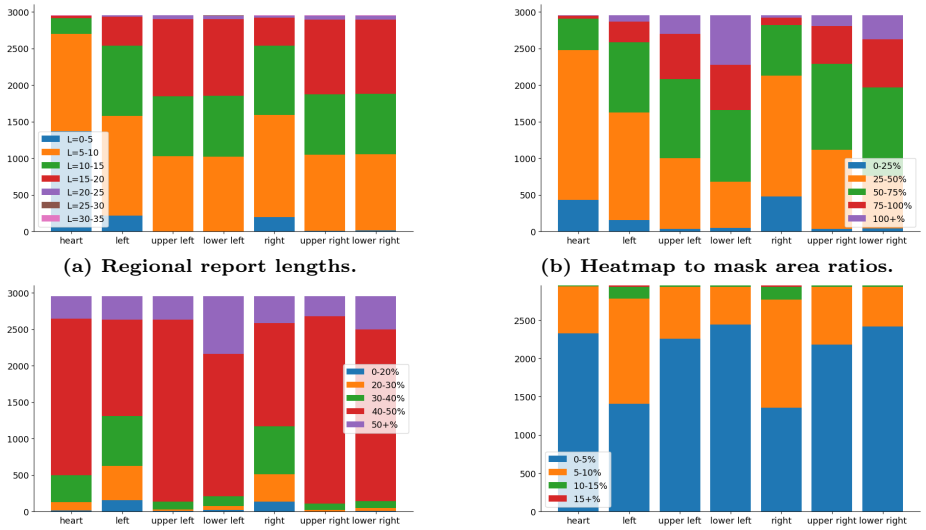
According to radiologists, their focus can be divided into seven key areas of CXR: heart, left, right, upper left, upper right, lower left, and lower right lungs. Therefore, we create anatomical masks for seven regions. Leveraging CXRs and gaze sequences from EGD [17] and REFLACX [1], we apply techniques from [34] to generate detailed masks for the heart and lungs, segmented into upper and lower regions. Finally, images with extreme brightness are filtered out.

3.2 Anatomical-aware Gaze Attention.

Given the gaze coordinates $G = \{g_1, g_2, \ldots, g_{|G|}\} \in \mathbb{N}^{|G|\times 2}$ of a CXR, our filtering process as the follows: For a report $T = \{s_1, s_2, ...s_{|T|}\}$, we identify all s_i that include keywords pertinent to the area of interest, select the latest end time and remove all gaze points after that timestamp. The list of keywords is described in Table 2. If transcript lacks keywords indicating anatomy, we use the entire gaze sequence. Finally, we filter out any gaze points that fall outside the segmentation masks corresponding to the anatomical areas of interest. The final gaze sequence of a i^{th} sample is represented in two forms: gaze sequence in a temporal order $G^i \in \mathbb{N}^{|G^i|\times 2} \subseteq G$, and gaze attention heatmap $A \in \mathbb{R}^{H\times W}$, which is created by creating gaze frequency map and applying Gaussian blurring as in [17].

3.3 Anatomical-aware Report

For every anatomical region, we associate it with a brief report. For instance, we link "the heart is normal" with the heart area. However, a report from REFLACX [1] or EGD [17] might only include certain anatomical regions. To address this, we use a template to generate reports for any missing anatomy. If a diagnosis for any region is absent after keyword filtering (Section 3.2), we



(a) Regional report lengths.

(b) Heatmap to mask area ratios.

(c) Heatmap to bounding box area ratios.

(d) Heatmap to full image size ratios per region.

Fig. 3: Dataset distributions for FG-CXR.

create a default sentence based on MIMIC-CXR annotations: **“the {area} is possibly normal”** for no findings, or **“the patient is possibly suffering from {findings} in the {area}”** for specific findings. For example, if a patient’s current report lacks information for the left lung area, we refer to MIMIC-CXR and find that the label for this patient is “no finding”. We then generate the sentence “the left lung is possibly normal” for this patient’s left lung.

3.4 Dataset Statistics

Our dataset contains 2,951 CXRs in total, with 20,657 pairs of {attention heatmap, report}. Figure 3 provides deeper insights into our dataset.

Reports are mostly concise. In Figure 3a, we observe that the majority of reports are concise, particularly those describing the heart, with most reports comprising fewer than 10 words. Reports on other anatomical regions tend to be longer, though a significant number still fall within the 5 to 10 words range. This indicates a prevalent practice among radiologists of using succinct sentences.

Radiologist’s attention versus segmentation mask. Figure 3b reveals that most gaze heatmaps typically cover a smaller area than the full anatomical segmentation mask. Notably, many heatmaps in the lower left and right lungs encompass a larger area (resulting in a ratio greater than 1). This can be attributed to the presence of dense gaze sequences that extensively cover a particular region, and this is further amplified with the Gaussian filtering process. Such a phenomenon is particularly evident in the lower right and lower left regions,

Table 3: Dataset splits for training, validation, and testing sets. From the data in Section 3, we create these splits for training, validating, and testing our method in Section 5.

Set	Number of Samples	Percentage
Training Set	2,074	70%
Validation Set	295	10%
Testing Set	582	20%
Total	2,951	**100%**

likely due to radiologists' meticulous examination of the diaphragm, which extends beyond the lower masks. Furthermore, the base of the left lung mask is typically smaller than that on the right side, attributed to occlusion by the heart.
Radiologist's attention versus bounding box. In Figure 3c, most heatmaps are confined to a portion of their bounding box, which is computed by taking the top left and bottom right corners of the non-zero heatmap values. A notable observation is that many gaze attention heatmaps utilize less than 20% of the bounding box area, especially in the left and right lungs.
Radiologist's attention versus the whole image. In Figure 3d, most heatmaps use little information from the whole image. This is true even for the left and right lungs, where one might expect a higher coverage area; however, most heatmaps occupy only about 10% of the image's total area.

For benchmarking in Section 5, we randomly split our dataset into 70% for training, 10% for validation, and 20% for testing. The number of samples is shown in Table 3.

3.5 How will our FG-CXR benefit the community?

We anticipate that the release of our FG-CXR will drive advancements in these promising areas:

- Gaze-Interpretable Report Generation: While report generation is a growing research topic, enhancing and evaluating report generation with explainability using radiologist gaze data remains relatively unexplored.
- General Medical Tasks: The FG-CXR dataset, enriched with segmentation masks for critical anatomical areas, serves as a valuable resource for developing and benchmarking Anatomical Segmentation algorithms [23, 47]. The reports of FG-CXR dataset, validated by experts, is also richer and more informative than the original REFLACX [1] and EGD [17] datasets, making it a potential benchmark for Radiology Report Generation [25, 55].

4 Methodology

In this section, we introduce a novel framework for Gaze-interpretable Report Generation. Given a CXR image $\mathcal{I}$, our goal is to produce gaze-based heatmaps $\mathcal{A}$

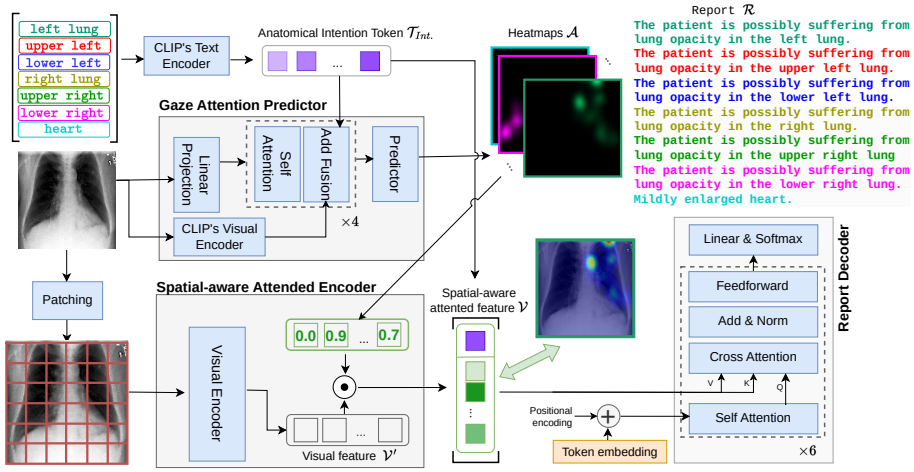


Fig. 4: The detailed architecture of our framework consisting of three key modules: Gaze Attention Predictor, Spatial-Aware Attended Encoder, and Report Decoder.

and generate a report $\mathcal{R}$ of the attended region $\mathcal{A}$. To ensure interpretability, the gaze attention $\mathcal{A}$ has to be closely aligned with expert observation and the report $\mathcal{R}$ has to be consistent with what gaze attention $\mathcal{A}$. Our Gen-XAI architecture, detailed in Figure 4, comprises three key modules: (i) Gaze Attention Predictor (GAP) to predict seven gaze-based heatmaps $\mathcal{A}$; (ii) Spatial-aware Attended Encoder (SAP) to generate attended features $\mathcal{V}$; and (iii) Report Decoder to generate diagnosis reports for seven anatomies.

4.1 Gaze Attention Predictor (GAP)

Given a CXR image $\mathcal{I}$ and Anatomical Intention Token $\mathcal{T}_{Int.}$, this module predicts the radiologist-like attention heatmap $\mathcal{A} \in [0,1]^{7\times H/16\times W/16\times 1}$. Inspired by [34], we adopt the idea of training a deep adapter on a pretrained CLIP [36] to predict heatmaps. Initially, we extract four intermediate features from the middle layers (i.e. $\{0,3,6,9\}$ according to [34]) of the CLIP's visual encoder from the $\mathcal{I}$. Subsequently, $\mathcal{I}$ is split into $H/16 * W/16$ patches with size 16×16 and projected into an embedding space $\mathcal{I}_e$ to prepare for the fusion stage via a Linear Projection layer. In the fusion stage, we fuse $\mathcal{I}_e$ with the extracted CLIP's visual features $\mathcal{V}$ and $\mathcal{T}_{Int.}$. Finally, a Predictor module, consisting of an MLP followed by a sigmoid activation, predicts seven gaze heatmaps for the seven parts of the lung. To create $\mathcal{T}_{Int.}$, we use CLIP's text encoder to extract seven textual features from: "`heart`", "`left`", "`right`", "`upper left`", "`upper right`", "`lower left`", and "`lower right`"; and we stack them into a tensor $\mathcal{T}_{Int.}$ of size $7 \times D$. This allows simultaneous prediction of all seven parts, rather than individually.

4.2 Spatial-aware Attended Encoder (SAP)

This module produces an attended feature $\mathcal{V}$ that contains information from its patch and neighbors, crucial for focusing on relevant areas while retaining essential spatial context. Given CXR's grayscale nature, distinguishing areas like the lung from the background requires understanding their spatial relation to adjacent patches. This spatial information only exists in the encoded feature [21,50]. Unlike previous works [18,34] that apply attention to pixel level that may remove vital details, our approach applies attention to the latent visual features $\mathcal{V}'$. The effectiveness is empirically proven and included in Section 5.3. Specifically, we first extract the patches' visual feature $\mathcal{V}' \in \mathbb{R}^{H/16 \times W/16 \times D}$ by using a Visual Encoder (i.e. CvT [50]). Then, we create the spatial-aware attended feature by performing element-wise multiplication between $\mathcal{A}$ and $'$ to create the reweighted feature $\mathcal{V}$. To further guide the model, we also concatenate the token of the current area of interest into the feature, for example concatenating the intention token of looking at the heart to the feature that is masked by the heart heatmap. Mathematically, we compute $\mathcal{V} \in \mathbb{R}^{7 \times (H/16 * W/16+1) \times D}$ with $\mathcal{V}(i) = [\mathcal{V}' \odot \mathcal{A}(i), \mathcal{T}_{Int.}(i)], \forall i \in [0, 6]$, where i indicates i^{th} region, $[\cdot]$ is concatenation and $\odot$ is the Hadamard product.

4.3 Report Decoder

We utilize GPT2 [36], an auto-regressive network for text generation, as our textual report decoder architecture. The token embedding is the embedding of previous tokens, for example, ‘‘`[BOS],the,patient,is, possibly,suffer,from`’’ to predict the next word ‘‘`lung`’’, where `[BOS]` is the beginning of sentence token. For every area i, we use $\mathcal{V}(i)$ as key (`K`) and value(`V`), and the output feature from self attention of token embedding as query (`Q`)of the cross-attention module. After predicting all sentences, we concatenate them to create the final report.

4.4 Learning Objective

We train our model with the training loss $\mathcal{L} = (1 + \lambda_c)\mathcal{L}_c + (1 + \lambda_h)\mathcal{L}_h$, where $\lambda_c, \mathcal{L}_c, \lambda_h, \mathcal{L}_h$ are the report penalty, cross-entropy loss for the generated report, heatmap penalty, and L_2 loss for predicted heatmap, respectively. To enhance the model's focus on predicting correct anatomy, we introduce two dynamic coefficients as penalties: gaze attention prediction (λ_h) and report generation (λ_c). During the heatmap prediction, we use Intersection over Union (IoU) with a threshold of 0.5 to identify instances where the model inaccurately focuses. Each incorrect prediction increases λ_h by 1. $\mathcal{A}_{gt}$ is the ground truth gaze attention map.

$$\lambda_h = \sum_i \mathbb{1}_{0.5}(\text{IoU}(\mathcal{A}(i), \mathcal{A}_{gt}(i))), \text{ where } \mathbb{1}_{0.5} = \begin{cases} 1 & \text{if } x \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

At the report generation, we want the model the model explicitly predict directions while minimizing incorrect directional words. Thus, if the model fails

Table 4: Performance comparison between our method and other SOTA methods on natural language generation (NLG) metrics for the full report generation task.

Methods	B1↑	B2↑	B3↑	B4↑	M↑	R↑	C↑	Div@2↑	R@4↓
R2Gen [4]	0.690	0.576	0.502	0.459	0.326	0.607	3.302	0.357	0.459
R2GenCMN [3]	0.688	0.572	0.513	0.472	0.326	0.612	3.355	0.437	0.174
CvT2DistilGPT2 [33]	0.708	0.647	0.585	0.552	0.352	0.618	2.936	0.486	0.149
RGRG [44]	0.715	0.598	0.583	0.550	0.351	0.532	3.300	0.624	0.075
$\mathcal{M}^2$ Transformer [6]	0.694	0.613	0.533	0.476	0.333	0.623	3.097	0.672	0.065
Ours	**0.729**	**0.658**	**0.606**	**0.561**	**0.386**	**0.692**	**4.026**	**0.854**	**0.055**

Table 5: Performance comparison between our method and other SOTA methods on clinical efficacy (CE) metrics.

Methods	P_{mic} ↑	R_{mic}↑	$F1_{mic}$↑	P_{mac}↑	R_{mac}↑	$F1_{mac}$↑	P_{ex}↑	R_{ex}↑	$F1_{ex}$↑
R2Gen [4]	0.415	0.249	0.319	0.259	0.135	0.151	0.370	0.237	0.254
R2GenCMN [3]	0.423	0.256	0.341	0.257	0.134	0.152	0.373	0.291	0.309
CvT2DistilGPT2 [33]	0.275	0.255	0.337	0.261	0.135	0.153	0.376	0.294	0.313
RGRG [44]	0.430	0.272	0.361	0.270	0.142	0.160	0.401	0.436	0.427
$\mathcal{M}^2$ Transformer [6]	0.467	0.429	0.459	0.283	0.171	0.197	0.436	0.461	0.440
Ours	**0.495**	**0.515**	**0.505**	**0.311**	**0.256**	**0.256**	**0.515**	**0.503**	**0.497**

Table 6: Performance comparison between our method and other SOTA methods for attention generation.

Methods	fgIoU↑	bgIoU↑	fwIoU↑	SSIM↑	PSNR↑	L1↓	L2↓
R2Gen [4]	15.87	56.03	45.08	0.35	9.12	0.840	0.200
R2GenCMN [3]	18.84	64.55	51.72	0.37	10.81	0.179	0.049
CvT2DistilGPT2 [33]	17.73	66.48	48.61	0.41	11.09	0.271	0.065
RGRG [44]	21.53	66.98	55.65	0.41	12.44	0.210	0.055
$\mathcal{M}^2$ Transformer [6]	23.19	69.06	60.22	0.45	14.51	0.120	0.031
Ours	**30.15**	**89.08**	**80.69**	**0.60**	**17.41**	**0.084**	**0.022**

to predict anatomical keywords or mentions the wrong direction, λ_c increases by 1. For every CXR, both λ_h and λ_c are initialized to 0, indicating that they are not accumulated across all samples in an epoch. The ablation study on the effect of penalty terms is included in Section 5.3.

5 Experiments

5.1 Implementation details

Architecture details. GAP comprises an FCN layer for an input patch size of 16×16 as the Linear Projection, 4 fusion layers, each with the Add Fusion block [34], and the Self-attention block with a hidden dimension of 240 and 6 attention heads. A BiomedCLIP [54] is used as CLIP and an MLP with 3 hidden layers of 256 neurons each as the Predictor. The Report Decoder is a GPT2 [36] initialized with DistillGPT2 [39] that has 12 heads, 6 layers, and a hidden dimension of 768. The SAP's Visual Encoder is a CvT [50] initialized with ImageNet [9] and $|\mathcal{V}| \in R^{768}$. We train Gen-XAI with a learning rate of $5e-5$, batch size of 32, $6,000$ iterations, and AdamW optimizer [26].

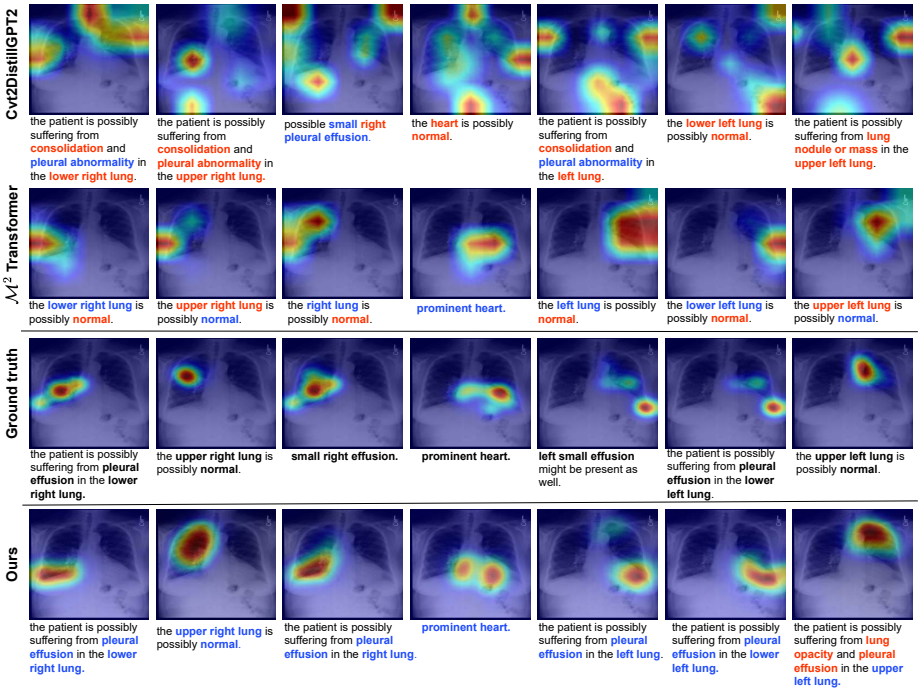


Fig. 5: Qualitative comparison. The blue text highlights consistent content with ground truth. The red text indicates incorrect content.

Metrics. Follow [34, 41, 43], we evaluate our method based on three criteria:

- Natural Language Generation (NLG) metrics: BLEU (B), METEOR (M), ROUGE-L (R), and CIDEr (C) for matching generated report with the reference report; Div@2 [41] and R@4 [52] for diversity generated reports because we observe that report generation model can suffer heavily from generating only one sentence for all samples.
- Clinical Efficacy (CE) metrics: we use all micro-, macro-, and example-based Precision, Recall, and F1 score described in [13] because NLG metrics alone are ill-suited for measuring clinical correctness [43].
- Attention Similarity metrics: we report all foreground IoU (fgIoU), background IoU (bgIoU), frequency-weighted IoU (fwIoU), Structural Similarity (SSIM), Peak signal-to-noise ratio (PSNR), L1, and L2. Note that, these metrics can also indicate interpretability because when a model's predicted areas of focus closely match those of expert radiologists, it suggests that the model's decision-making process is more understandable and interpretable.

Baselines. We compare Gen-XAI with state-of-the-art methods: R2GenCMN [3], R2Gen [4], CvT2DistilGPT2 [33], $\mathcal{M}^2$ Transformer [6], and RGRG [44] on FG-CXR. For each method, we maintain the default hyperparameters as specified by the authors and train all models on our dataset. For previous works, we

Table 7: Ablation study on applying attention ($\mathcal{T}_{Int}$) to pixels vs. features.

Settings	$\mathcal{T}_{Int}$	Attention			NLG			CE		
		fwIoU↑	**PSNR↑**	**L1↓**	**B4↑**	**C↑**	**Div@2↑**	$\mathbf{P}_{ex}$↑	$\mathbf{R}_{ex}$↑	$\mathbf{F1}_{ex}$↑
On pixel	✗	62.07	13.50	0.122	0.428	3.004	0.678	0.400	0.416	0.410
	✓	73.35	15.97	0.094	0.464	3.300	0.734	0.434	0.453	0.447
On feature	✗	79.11	17.24	0.085	0.545	3.947	0.837	0.505	0.498	0.487
	✓	80.69	17.41	0.084	0.561	4.026	0.854	0.515	0.503	0.497

use the attention scores of the generated sentences as predicted attention. Since RGRG employs a single vector to represent each region, we use its predicted bounding box for evaluation.

5.2 Experimental results

Quantitative results. Tables 4 to 6 demonstrates the effectiveness of our interpretable approach, which outperforms other methods across all criteria. For example, Gen-XAI has a high advantage in attention prediction and outshines the runner-up [6] by +20.47 in fwIoU. The improvement in attention similarity is understandable as our method is explicitly constrained by the heatmap loss while other methods are not designed for Gaze-Interpretable Report Generation. However, Gen-XAI also outperforms other black box methods, designed specifically for radiology report generation. For example, in NLG, Gen-XAI significantly surpasses other models by a large margin, i.e. +0.671 on CIDEr metric compared to the runner-up [3], and +0.182 on Div@2 metric compared to the runner-up [6]. On example-based CE metrics, our model also achieves a higher score of 0.497, +0.057 on $F1_{ex}$ respectively compared to the runner-up [6]. One of possible reason is that previous works are not supervised by radiologist's gaze attention, and thus they fail to learn and use incorrect visual information. This will be further confirmed in Figure 5, where looking at the wrong location causes the model to fail in producing reliable diagnosis, and in Section 5.3, where our model also encounters the same issue when trained with a traditional attention mechanism.

Qualitative results. Figure 5 compares our Gen-XAI with leading methods, showcasing superior performance in generating precise attention heatmaps and diagnosis reports. CvT2DistilGPT2 produces unreliable and mostly incorrect reports due to inaccurate focus, despite occasional recognition of pleural effusion. On the other hand, the $\mathcal{M}^2$ Transformer often misidentifies lungs as normal, although accurately diagnosing the heart. This highlights the effectiveness of our approach and the crucial role of precise anatomical focus in addressing the Gaze-Interpretable Report Generation challenge.

Table 8: Ablation study on the penalty effects.

Settings	Attention			NLG			CE		
	fwIoU↑	PSNR↑	L1↓	B4↑	C↑	Div@2↑	P_{ex}↑	R_{ex}↑	$F1_{ex}$↑
w/o. penalty	78.12	16.58	0.088	0.515	3.763	0.849	0.490	0.479	0.478
+ λ_c	77.41	16.74	0.088	0.534	3.871	0.837	0.500	0.493	0.483
+ λ_h	79.74	17.07	0.085	0.524	3.834	0.837	0.500	0.488	0.478
+ $\lambda_c + \lambda_h$	80.69	17.41	0.084	0.561	4.026	0.854	0.515	0.503	0.497

Table 9: Ablation study on training with the proposed anatomical gaze attention vs. anatomical segmentation vs. traditional attention.

Settings	Attention			NLG			CE		
	fwIoU↑	PSNR↑	L1↓	B4↑	C↑	Div@2↑	P_{ex}↑	R_{ex}↑	$F1_{ex}$↑
Traditional attention	69.71	14.59	0.148	0.395	3.113	0.656	0.399	0.411	0.406
Anatomical segmentation	79.95	17.11	0.090	0.551	3.998	0.849	0.485	0.490	0.488
Anatomical gaze attention	80.69	17.41	0.084	0.561	4.026	0.854	0.515	0.503	0.497

5.3 Ablation Study

Applying gaze attention on pixel vs. features. In Section 4, we apply the predicted gaze attention on the features based on the intuition that the encoder may extract important context information to represent a patch feature besides the patch pixels. For example, the shape or spatial information can be in the feature. This ablation alters only the attention choice, keeping penalty terms and other components as initially proposed. Therefore, we design an ablation study: Instead of applying the gaze attention on the feature, we apply it to the image input, then feed the masked image into the encoder again. Every other setting is kept the same. The results are shown in Table 7. Indeed, the findings indicate that incorporating gaze attention directly into the input detracts from model performance. As mentioned in Section 4, this approach can obscure spatial details, leading to confusion. For instance, a heatmap centered on the left lung may not clarify enough whether it targets the left or right side due to uniform coloration. This issue is amplified when training the model without an intention token (w/o. IT), as shown in our table's first row. On the other hand, attending to the latent feature improves the performance, and applying the intention token can slightly boost the performance.

Anatomical gaze attention vs. anatomical segmentation vs. traditional attention. Interpretable model is often mistakenly thought to be harmful to the performance [37]. To demonstrate that this is not the case for our proposed model. We design an ablation study with two more settings:

– *Traditional attention.* We remove the Gaze Attention Predictor (GAP) module and instead use a simple self-attention module to flexibly weigh the impor-

tance of every patch feature. In other words, we let the model automatically decide the importance of every patch.
- *Anatomical segmentation.* Instead of supervising our model on gaze attention ground truth, we supervise the GAP with our anatomical segmentation masks. Then we mask out patches that are not in the predicted mask. In other words, we let the GAP module be an anatomical mask predictor, and a patch is important, i.e. its weight is 1.0 if it is inside the anatomy of interest.

Table 9 shows that our proposed anatomical gaze attention supervision is effective and outperforms other settings. For the traditional attention setting, the black-box and unconditional training pipeline causes the model to not know where to look, i.e. low scores on Attention criteria, and hence, it fails to give satisfaction diagnosis, i.e. low NLG and CE scores. One of the possible reasons for this is because self self-attention mechanism is well-known for its data-hungry nature [19]. On the other hand, training the GAP module with segmentation masks slightly decreases the performance. One possible reason is that we let the model use too much information, which can confuse the model in some cases. The gain from correctly weighing important patches further confirms our hypothesis in Section 1. Moreover, this suggests that our framework can also be used for segmentation prediction.

Effectiveness of penalty terms. The intuition behind the penalty terms is simple, yet effective. We design an ablation study: we train the model without penalty terms to demonstrate the effect of every penalty. As a result, we find that our penalties based on the idea of looking at the correct anatomy are beneficial to the model, as shown in Table 8.

6 Conclusion

In this work, we have introduced FG-CXR, a curated dataset for gaze interpretable radiology report generation. Our dataset contains CXR images with aligned gaze sequence, gaze attention heatmap, and the reports associated with seven anatomical parts of the lung. We then presented a novel method for generating descriptive reports of chest X-ray images, using heatmaps based on radiologist annotations to focus the model's attention and reduce the likelihood of misinterpreting irrelevant regions. Our main contribution is the successful application of a radiologist-informed attention mechanism that guides a generative model, thereby enhancing the accuracy, reliability, and interpretability of automated CXR report generation. We hope that the release of our dataset will advance more research on interpretable report generation.

Acknowledgments. This material is based upon work supported by the National Science Foundation (NSF) under Award No OIA-1946391, NSF 2223793 EFRI BRAID, National Institutes of Health (NIH) 1R01CA277739-01.

References

1. Bigolin Lanfredi, R., Zhang, M., et al.: Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. Scientific data (2022) 2, 4, 5, 7
2. Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis (2020) 4
3. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022) 10, 11, 12
4. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020) 10, 11
5. Coffman, E., Clark, R., Bui, N.T., Pham, T.T., Kegley, B., Powell, J.G., Zhao, J., Le, N.: Cattleface-rgbt: Rgb-t cattle facial landmark benchmark. arXiv preprint arXiv:2406.03431 (2024) 2
6. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-Memory Transformer for Image Captioning. In: CVPR (2020) 10, 11, 12
7. Datta, S., Roberts, K.: A dataset of chest x-ray reports annotated with spatial role labeling annotations. Data in Brief (2020) 4
8. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association (2016) 4
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 10
10. Filice, R.W., Stein, A., et al.: Crowdsourcing pneumothorax annotations using machine learning annotations on the nih chest x-ray dataset. Journal of digital imaging (2020) 3, 4
11. Geis, J.R., Brady, A.P., Wu, C.C., Spencer, J., Ranschaert, E., Jaremko, J.L., Langer, S.G., Borondy Kitts, A., Birch, J., Shields, W.F., et al.: Ethics of artificial intelligence in radiology: summary of the joint european and north american multisociety statement. Radiology **293**(2), 436–440 (2019) 2
12. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018) 2
13. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI (2019) 4, 11
14. Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X., Thoma, G.: Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery (2014) 4
15. Jing, B., Wang, Z., Xing, E.: Show, describe and conclude: On exploiting the structure information of chest x-ray reports. arXiv preprint arXiv:2004.12274 (2020) 3
16. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., et al.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data (2019) 4
17. Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J., Tong, M., Sharma, A., Abedin, S., Beymer, D., Mukherjee, V., Krupinski, E., et al.: Eye gaze data for chest x-rays. PhysioNet (2020) 2, 4, 5, 7

18. Kashyap, S., Karargyris, A., Wu, J., Gur, Y., Sharma, A., Wong, K.C., Moradi, M., Syeda-Mahmood, T.: Looking in the right place for anomalies: Explainable ai through automatic location learning. In: ISBI (2020) 9
19. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM computing surveys (CSUR) **54**(10s), 1–41 (2022) 14
20. Kim, B., Wattenberg, M., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: ICML (2018) 3
21. Le, M.Q., Graikos, A., Yellapragada, S., Gupta, R., Saltz, J., Samaras, D.: ∞-brush: Controllable large image synthesis with diffusion models in infinite dimensions. arXiv preprint arXiv:2407.14709 (2024) 2, 9
22. Le, N., Pham, T., Do, T., Tjiputra, E., Tran, Q.D., Nguyen, A.: Music-driven group choreography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8673–8682 (2023) 2
23. Lei, B., Huang, S., et al.: Self-co-attention neural network for anatomy segmentation in whole breast ultrasound. Medical image analysis (2020) 7
24. Li, Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. Advances in neural information processing systems (2018) 3
25. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: CVPR (2021) 4, 7
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 10
27. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019) 2
28. Nauta, M., Schlötterer, J., van Keulen, M., Seifert, C.: Pip-net: Patch-based intuitive prototypes for interpretable image classification. In: CVPR (2023) 3
29. Nguyen, T.P., Pham, T.T., Nguyen, T., Le, H., Nguyen, D., Lam, H., Nguyen, P., Fowler, J., Tran, M.T., Le, N.: Embryosformer: Deformable transformer and collaborative encoding-decoding for embryos stage development classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1981–1990 (2023) 2
30. Nguyen, V.D., Khaldi, K., Nguyen, D., Mantini, P., Shah, S.: Contrastive viewpoint-aware shape learning for long-term person re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1041–1049 (2024) 2
31. Nguyen, V.D., Mantini, P., Shah, S.K.: Occluded cloth-changing person re-identification via occlusion-aware appearance and shape reasoning. In: 2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–8. IEEE (2024) 2
32. Nguyen, V.D., Mirza, S., Zakeri, A., Gupta, A., Khaldi, K., Aloui, R., Mantini, P., Shah, S.K., Merchant, F.: Tackling domain shifts in person re-identification: A survey and analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4149–4159 (2024) 2
33. Nicolson, A., Dowling, J., Koopman, B.: Improving chest X-ray report generation by leveraging warm starting. Artificial Intelligence in Medicine (2023) 10, 11
34. Pham, T.T., Brecheisen, J., Nguyen, A., Nguyen, H., Le, N.: I-ai: A controllable & interpretable ai system for decoding radiologists' intense focus for accurate cxr diagnoses. In: WACV (2024) 2, 3, 5, 8, 9, 10, 11
35. Pham, T.T., Do, T., Le, N., Le, N., Nguyen, H., Tjiputra, E., Tran, Q., Nguyen, A.: Style transfer for 2d talking head generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7500–7509 (2024) 2

36. Radford, A., Wu, J., et al.: Language models are unsupervised multitask learners. OpenAI blog (2019) 8, 9, 10
37. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence (2019) 2, 13
38. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys (2022) 3
39. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019) 10
40. Selvaraju, R.R., Cogswell, M., et al.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: CVPR (2017) 3
41. Shetty, R., Rohrbach, M., Anne Hendricks, L., Fritz, M., Schiele, B.: Speaking the same language: Matching machine to human captions by adversarial training. In: ICCV (2017) 11
42. Shih, G., Wu, C.C., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence (2019) 3, 4
43. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: CVPR (2023) 2, 3, 4, 11
44. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: CVPR (2023) 10, 11
45. Team, P.P., Gohagan, J.K., Prorok, P.C., Hayes, R.B., Kramer, B.S.: The prostate, lung, colorectal and ovarian (plco) cancer screening trial of the national cancer institute: history, organization, and status. Controlled clinical trials (2000) 4
46. Tran, M.T., Nguyen, T.V., Hoang, T.H., Le, T.N., Nguyen, K.T., Dinh, D.T., Nguyen, T.A., Nguyen, H.D., Hoang, X.N., Nguyen, T.T., et al.: itask-intelligent traffic analysis software kit. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 612–613 (2020) 2
47. Ullah, I., Ali, F., Shah, B., El-Sappagh, S., Abuhmed, T., Park, S.H.: A deep learning based dual encoder–decoder framework for anatomical structure segmentation in chest x-ray images. Scientific Reports (2023) 7
48. Vo, K., Pham, T.T., Yamazaki, K., Tran, M., Le, N.: Dna: Deformable neural articulations network for template-free dynamic 3d human reconstruction from monocular rgb-d video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3676–3685 (2023) 2
49. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR (2017) 4
50. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: ICCV (2021) 9, 10
51. Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Mitchell, W., Kashyap, S., Giovannini, A., et al.: Chest imagenome dataset (version 1.0. 0). PhysioNet (2021) 2, 3, 4
52. Xiong, Y., Dai, B., Lin, D.: Move forward and tell: A progressive generator of video descriptions. In: ECCV (2018) 11
53. You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: MICCAI (2021) 3

54. Zhang, S., Xu, Y., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023) 10
55. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: AAAI (2020) 7